DOCUMENT RESUME

ED 054 193 TM 000 757

AUTHOR Zedeck, Sheldon; Baker, Henry T.
TITLE Evaluation of Behavioral Expectation Scales.
PUB DATE May 71
NOTE 8p.; Paper presented at the Annual Meeting of the Midwestern Psychological Association, Detroit, Michigan, May 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Attendance, Behavioral Objectives, Behavioral Science Research, *Behavior Rating Scales, Correlation, *Evaluation, *Expectation, *Nurses, Performance Criteria, *Supervisors, Tenure, Test Reliability
IDENTIFIERS Behavioral Expectation Scales, BES

ABSTRACT

Behavioral Expectation Scales developed by Smith and Kendall were evaluated. Results indicated slight interrater reliability between Head Nurses and Supervisors, moderate dependence among five performance dimensions, and correlation between two scales and tenure. Results are discussed in terms of procedural problems, critical incident problems, and perspective of raters. (Author)

EVALUATION OF BEHAVIORAL EXPECTATION SCALES[1]

Sheldon Zedeck and Henry T. Baker

University of California, Berkeley

Recently, several investigators (Campbell, Dunnette, Lawler, & Weick, 1970; Fogli, Hulin, & Blood, 1971; Landy & Guion, 1970; Maas, 1965) have used or recommended using behaviorally anchored rating scales which are constructed using a procedure reported by Smith and Kendall (1963). Briefly, the Smith and Kendall procedure is an iterative technique which involves the development of dimensions, scales, and items of performance criteria by independent groups. There are several advantages of behavioral expectation scales (BES): (1) groups with work experiences similar to those who eventually use the scales participate in the construction of the scales; (2) behavioral incidents are used as anchor points on each scale; (3) the terminology used on the job is retained in the anchors; (4) relatively independent scales with high scale reliabilities are obtained; and (5) in actual use, raters document their ratings with specific incidents.

Smith and Kendall (1963) used the above procedure to develop scales to be used for evaluation of staff nurses. Four groups of head nurses, from different hospitals, participated in the study. Final results yielded five dimensions on which staff nurses could be evaluated: knowledge and judgment, conscientiousness, skill in human relations, organizational ability, and observational ability. Scale reliabilities of the items were .97 or better. However, very little is known about the construct validity of the scales. Thus, the purpose of the present study was to evaluate BES with respect to interrater reliability, the independence of five dimensions, the relation of the scales to other criteria,

---

[1] Paper presented at Midwestern Psychological Association, Detroit, May 1971.

and the practical usefulness of the scales.

Method

Subjects

The research site was a public, non-profit, 165-bed hospital in Northern California. Two nursing supervisory levels, Head Nurses (HN) (N = 9) and Supervisors (SUP) (N = 5), participated in the performance evaluation of 98 staff registered nurses (RN). The span of control for HNs and SUPs ranged from four to seventeen, with a given HN and SUP pair not necessarily having a common set of RNs to evaluate. For each dimension correlations were obtained between two evaluation distributions, one provided by HNs and one provided by SUPs, for the sample of RNs who were common to both rater samples. Depending on raters involved and dimension evaluated, and because of missing data and incorrect use of BES, final intercorrelation sample sizes varied from 71 to 92.

Procedure

The BES developed by Smith and Kendall (1963) was used for performance evaluations. The appraisal procedure required that several incidents be noted for each dimension, that each incident be assigned a value with the behavioral anchors serving as a frame of reference, and finally, that the average of the values of the incidents for a given dimension be used as a summary value (Tate, 1964).

Each HN and SUP was given a brief training session on the use of BES. Each rater was asked to record incidents on each dimension for each nurse for whom she was responsible. The incidents were to reflect past performance of the RNs or performance as observed in a following two-month period. A restriction of at least one incident and (so that time spent on evaluations would not become excessive) no more than five incidents for each dimension was imposed. (The raters were given time off from regular duties to work on the BES evaluations.)

Tenure and absentee data were collected and correlated with performance evaluations. Absentee information was recorded dichotomously, an RN being either a chronic absentee or a non-absentee. Absentee information was obtained from a list the hospital administration had prepared, while the present research was in operation, because of concern about increasing chronic absenteeism.

In addition, six months after the ratings were completed, all incidents contributed by the raters and the original anchor items on the BES were randomized and presented to five judges (nurses) for revaluation. The five judges were to examine each incident, assign it to one of the five BES dimensions, and also to assign to it a value from 0.00 to 2.00, which was the scale range Smith and Kendall (1963) used. This procedure provided an examination of the appropriateness of the items for a dimension and the interrater reliability of item values.

Results

Results are shown in Table 1. The correlation diagonal in parentheses, representing interrater agreement, provided evidence that HNs and SUPs were in significant agreement in rating RNs on all five dimensions. Agreement between the two rater samples was best for human relations skill and poorest for conscientiousness and observational ability.

The degree of independence of the dimensions for HNs and SUPs is shown in the solid triangles. For both rater samples the intercorrelations between dimensions were significant and high. The amount of dependence ranged from .38 to .62 for HNs and from .69 to .82 for SUPs.

Table 1 also shows the relationship of absenteeism and of tenure to each of the performance dimensions for both groups of raters. Tenure was significantly correlated with knowledge and judgment and with conscientiousness as appraised by HNs. None of the correlations between absenteeism or tenure and the performance dimensions as appraised by SUPs were significant. The relationship between tenure and absenteeism was .37 ($N = 95$, $p < .01$).

Discussion

The major purpose of the present study was to evaluate the Smith and Kendall (1963) BES in a field situation in terms of construct validity. As a result, interesting problems pertaining to BES in particular and to performance appraisal in general surfaced.

The moderate degree of interrater reliability caused us to reassess our expectation that the two supervisory levels would agree on appraisals. In essence, BES scales were developed by supervisory personnel, retained nursing terminology, and provided meaningful dimensions. However, the two present supervisory groups were not equivalent to each other in their opportunity to evaluate staff nurses. HNs supervise wards and have direct contact with RNs, thereby placing them in an adequate position to observe and evaluate the RNs. On the other hand, SUPs are involved in more administrative functions, coordinating the activities in various areas of the hospital. As a result, SUPs have less opportunity to observe and evaluate RNs.

Hence, there are several reasons for the moderate amount of interrater reliability obtained in the present study. First, SUPs and HNs are two supervisory levels, each with different functions. Second, SUPs and HNs do not have equivalent opportunity to observe and evaluate RNs. Third, an hypothesis that SUPs and HNs emphasize similar behaviors has not been tested. If their different perspectives also cause the two groups to expect or value different behaviors, then the moderate interrater reliability is not surprising. In fact, geographical, demographic, or organizational differences between the nurses in this study and those in the Smith and Kendall (1963) study may be related to variation in perspective and expectations and, consequently, may account for the present results.

With respect to independence of dimensions, Smith and Kendall (1963) found five independent traits, independent in the sense that items and dimensions retained were meaningful and distinguishable for the original research samples.

The generalizability of the Smith and Kendall items is confirmed by the present judges, who exhibited high dimension and value agreement. That is, the five judges correctly categorized the original BES items. (Per cent correct categorization for the judges ranged from 76% to 93%.)

However, the high interdimension correlations can be explained by considerable within-rater halo and method bias. The incidents provided by the present raters were vague, and there was only moderate agreement between judges as to their appropriateness for the indicated dimensions. (Per cent correct categorization for the judges ranged from 53% to 59%.)

Thus, the problem becomes one of utility of the specific BES; are the scales appropriate for one or more groups of raters? Several attempts were undertaken to investigate this problem. First, moderate interrater reliability but high dimension intercorrelation, indicating poor construct validity, ordinarily would limit the usefulness of new scales. However, given the extensive development procedure and the fact that we are dealing with perceptions of behavior, additional relationships were examined. The significant correlations between tenure and knowledge and judgment, and between tenure and conscientiousness, indicated that the relatively subjective performance BES criteria have some meaningful variance in common with objective criteria. Since the HNs' performance evaluations correlated with tenure, support is obtained for HNs as the appropriate raters. Second, support for HNs as appropriate raters pertains to amount of contact with RNs. Organizational structure prescribes that HNs have more contact with RNs, and consequently have more opportunity to observe behavior. Third, examination of the specific incidents cited and their accompanying values indicated that, in general, the items for a given scale provided by HNs were more consistent, whereas the items for a given scale by SUPs were more variable, and often the two items provided by SUPs were at opposite poles. Fourth, the intercorrelations between dimensions were less for HNs than for SUPs.

As a result, the authors conclude that the scales are more appropriate for HNs than for SUPs in the present study. The authors suggest examining the procedure as used in the present study as opposed to the one suggested by Smith and Kendall (1963). The present raters provided one to five incidents, which were subsequently averaged. In contrast, Smith and Kendall suggested that the rater provide a summary rating and then document the rating. The latter procedure eliminates variance within ratings for a given rater on a given scale applied to a given ratee. However, the question arises as to whether making a summary rating will dictate the type of documentations and, in effect, will provide an accurate description of the ratee's behavior. Both of the above procedures should be tried and the results compared.

A final recommendation is that considerable training be given to the raters in the use of the scales, particularly in an attempt to get support from the raters. The present raters were given approximately a half-hour training session in the use of BES. During the two-month observation and rating period several raters complained that the procedure was too time-consuming and difficult. However, when the research was completed, the raters were debriefed and given examples of "good" contributed incidents and "poor" contributed incidents. At that time some raters indicated that they had not known exactly what kind of incidents could be considered "good," which suggested that if they had received additional and more comprehensive training, results would have more accurately reflected the actual situation. In addition, most raters indicated that they preferred the BES format to the Likert-type rating scale which they had used in the past.

References

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. Managerial behavior, performance, and effectiveness. New York: McGraw-Hill, 1970.

Fogli, L., Hulin, C. L., & Blood, M. R. Development of first-level behavioral job criteria. Journal of Applied Psychology, 1971, 55, 3-8.

Landy, F. J., & Guion, R. M. Development of scales for the measurement of work motivation. Organizational Behavior and Human Performance, 1970, 5, 93-103.

Maas, J. B. Patterned scaled expectation interview: Reliability studies on a new technique. Journal of Applied Psychology, 1965, 49, 431-433.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Tate, B. Test of a nursing performance evaluation instrument. New York: National League for Nursing, 1964.

Table 1

Intercorrelations between Head Nurses and Supervisors on BES

		Head Nurse					Supervisor				
		K & J	CON	HR	ORG	OBS	K & J	CON	HR	ORG	OBS
Head Nurse	K & J										
	CON	.61*									
	HR	.38*	.54*								
	ORG	.50*	.57*	.45*							
	OBS	.61*	.62*	.56*	.54*						
Supervisor	K & J	(.47*)	.38*	.30*	.25**	.20					
	CON	.37*	(.25**)	.31*	.22**	.14	.72*				
	HR	.15	.31*	(.53*)	.07	.06	.49*	.62*			
	ORG	.53*	.43*	.52*	(.45*)	.35*	.62*	.52*	.60*		
	OBS	.49*	.32*	.31*	.28**	(.24**)	.82*	.73*	.57*	.65*	
	Tenure	.23**	.23**	.02	.06	.15	.13	.11	-.09	-.11	.08
	Absenteeism	.20	.17	.03	-.03	.05	.16	.03	-.01	.03	.10

NOTE.— K & J: Knowledge and Judgment
CON: Conscientiousness
HR: Skill in human relationships
ORG: Organizational ability
OBS: Observational ability

** $p < .05$
* $p < .01$